Data mining based cyber-attack detection

Tianfield, Huaglory

Published in:
System Simulation Technology

Publication date:
2017

Document Version
Author accepted manuscript

Data Mining Based Cyber-Attack Detection

Huaglory Tianfield

Glasgow Caledonian University, Glasgow G4 0BA, UK

Abstract: Detecting cyber-attacks undoubtedly has become a big data problem. This paper presents a tutorial on data mining based cyber-attack detection. First, a data driven defence framework is presented in terms of cyber security situational awareness. Then, the process of data mining based cyber-attack detection is discussed. Next, a multi-loop learning architecture is presented for data mining based cyber-attack detection. Finally, common data mining techniques for cyber-attack detection are discussed.

Keywords: big data analytics, cyber-attack detection, cyber security; cyber situational awareness; data mining; pattern mining; machine learning.

1. Introduction

The domain of cyber security is inherently a dynamically changing one. Newer attacks, such as multi-stage exploits and zero-day attacks, can be significantly more diverse than old attacks in terms of technical implementation as well as the underlying methods themselves in the ongoing wisdom race between attackers and defenders.

As cyber-attacks have evolved and grown in sophistication, cyber-attack detection techniques have also become much more sophisticated, by monitoring an ever increasing amount of diverse heterogeneous security event sources. [1] presents an overview upon various techniques under different detection models, namely misuse detection and anomaly detection, respectively, and also introduces degree of attack guilt as a way of characterizing intrusion detection activities.

Cyber security basically is about the capability to defence the information infrastructure from cyber-attacks. Given the trend that cyber-attacks get increasingly persistent, advanced, and stealthy, cyber defence has to rely upon systematic and intelligence oriented methodologies. Above all, getting a grip of the data view is essential. Across different cyber control domains and over a practical period of time, there are really voluminous and diverse sources of security information data. The ability to exploit and comprehend cyber security data enables or otherwise restricts the capability of cyber defence.

For instance, to tackle challenges like zero-day attacks, the only way ahead is to rely upon the development of a holistic, robust data driven approach.

Cyber-attack detection involves analysis of big data. For instance, host log event data can accumulate increasingly large. Such large volumes of data are overwhelming and a first challenge may simply be to store the data. With such big data issues at this scale, if analytical techniques cannot be taken use of effectively, false alarms are especially problematic.

Furthermore, to deal with the increased information security threats in large scale networks, many kinds of security devices have been used. These devices produce lots of security events. It can be very difficult to correlate events over such large amounts of data. Both the security alerting devices and alert messages can be heterogeneous in nature. The heterogeneity within a myriad of security technologies and systems which do not integrate well would cause difficulties for correlating the events.

In order to mitigate or prevent attacks, awareness of an attack is essential to being able to react and defend against attackers. Cyber defences can be further improved by utilizing security analytics to look for hidden attack patterns and trends. This is apparently the motivation for applying data mining for cyber security.

Data mining is the process of extracting useful and previously unnoticed models or patterns from large data stores [2]. The goal of data mining process is to discover patterns that are hidden among the huge sets of data. Thus, data mining is also taken as knowledge discovery.

[3] presents an overview on different data mining and machine learning algorithms being applied to misuse and anomaly detection. [4] presents a distributed architecture for data mining based intrusion detection that exploits the advantages of both misuse and anomaly detection strategies.

This paper presents a tutorial on data mining based cyber-attack detection. The remainder of the paper is arranged as follows. Section 2 presents a holistic data driven defence framework in terms of cyber security situational awareness. Section 3 discusses process of data mining based cyber-attack detection, especially from data processing, through data analysis, to data visualisation. Section 4 presents a multi-loop learning architecture for data mining based cyber-attack detection which combines detection models and learning modes. Section 5 discusses common data mining techniques for cyber-attack detection, including classification, clustering and association rule mining. Finally, Section 6 gives an outlook.

2. Data Driven Framework of Cyber Defence

The main idea of situational awareness in cyber domain is to analyse the surroundings in information infrastructure and to create certain events and visualizations for the purpose of efficient and fast decision-making. In simple words, cyber security situational awareness (CSAW) can be described as the situational awareness applied for cyber security in an information infrastructure.

From a systemic data point of view, cyber security situational awareness is handling data, correlating events, transforming data, discovering patterns, and inferring contexts and evidences, as illustrated in Figure 1. Through applying appropriate mechanisms of assessment, evaluation, inference, and so forth, CSAW forms perception and understanding of the situation and changes.

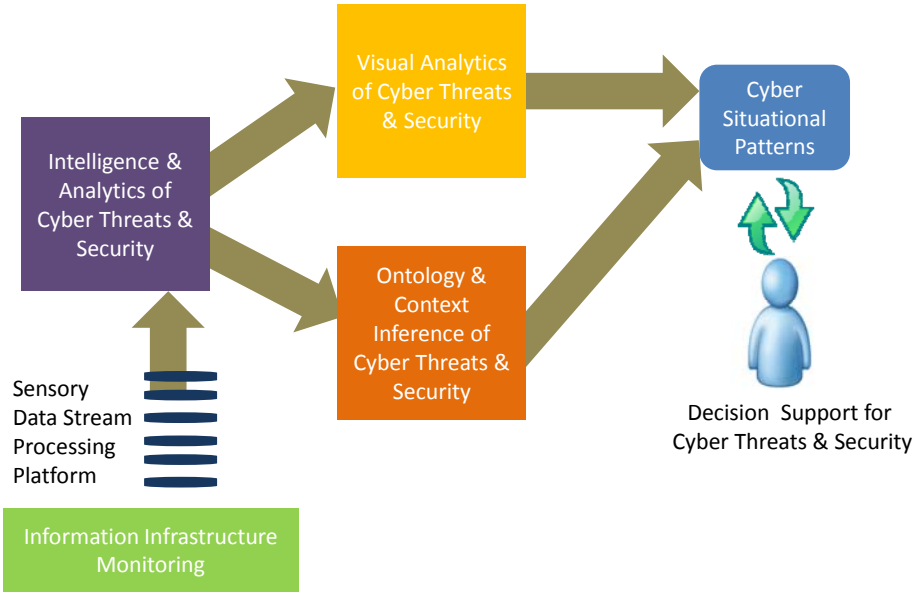

Figure 1. Cyber security situational awareness

Perception involves evidence gathering of the situations in the information infrastructure. Perception is to get the knowledge of the elements in the networked environment such as alerts reported by intrusion detection systems (IDS), firewall logs, scan reports, as well as the time they occurred.

Comprehension involves the analysis of the evidences to deduce the exact threat level, type of attack and associated or interdependent risks. Comprehension utilizes a set of relevant techniques and procedures to analyse, synthesize, correlate and aggregate pieces of evidence data perceived in the information infrastructure.

CSAW involves the perception of attacks and attack tracks, the comprehension of attack patterns and correlations, and the projection of what will happen in the near future in terms of impact and threat levels towards the information infrastructure [5].

CSAW is a process of data transformation and evidence refinement and valuation. The process of CSAW essentially corresponds to the lifecycle that security data should undergo [6], as illustrated in Figure 2. Along the lifecycle, data take different forms, ranging from the start at the raw sensor data, through cleansed data, fused data, correlated data, perceived events, and formulated contexts, and ending at the situational patterns. The upstream of the security data lifecycle is mainly concerned with data pre-processing, distributed data stores, and data fusion, while the downstream of the security data lifecycle is mainly concerned with event processing, situational assessment and modelling, sequential pattern mining and pattern analysis, context inference and management, and situational visualization.

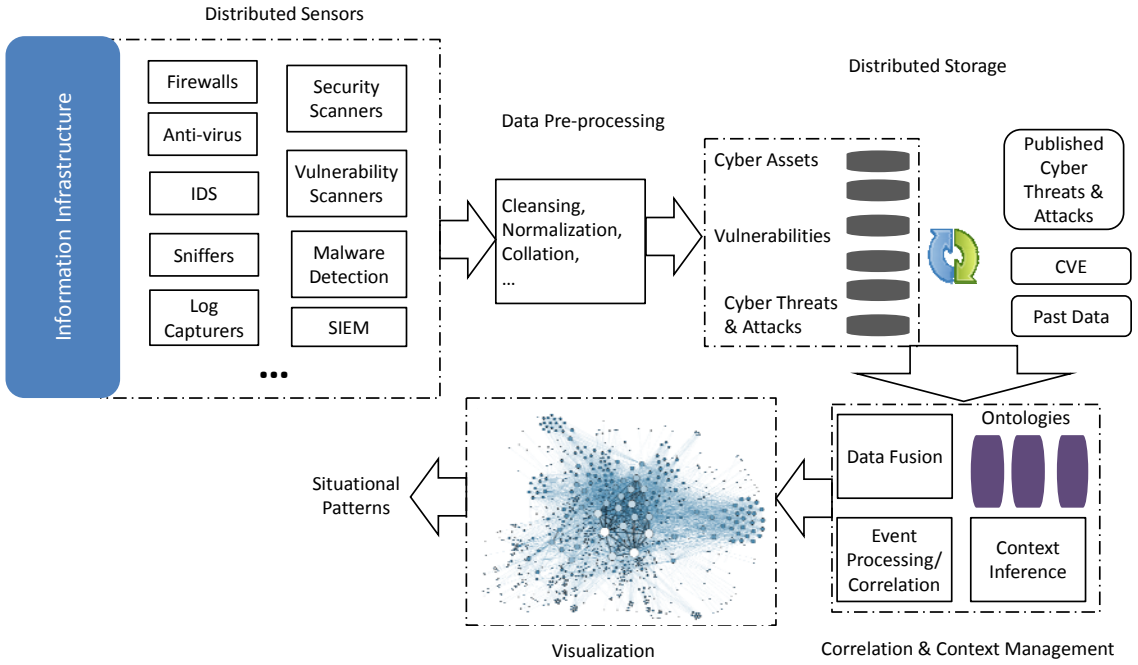

Figure 2. Data lifecycle framework of cyber security situational awareness
CVE: Common Vulnerabilities and Exposures; IDS: Intrusion Detection Systems; SIEM: Security Information and Event Management

As seen in Figure 2, from security data capturing to the cyber security situation model building, it is an integrated process. To exploit higher level values out of the security data, CSAW will undergo a multi-level analysis process. The information flow from security sensors to situational patterns forms an information value chain, thus a multi-level analysis framework, to achieve CSAW.

Along the security data lifecycle come the stages through which data are acquired, collected, processed, correlated and extracted for higher level values. Within each stage, the systems, techniques and toolkits are related to the major functionalities correspondingly needed. Moreover, CSAW is intrinsically born as a distributed data processing infrastructure.

CSAW starts at security monitoring. Security monitoring is about acquiring the ongoing phenomenon of computer or network system in which data continuously change. Through the instruments of a security operation centre, CSAW can gather network, system and application logs and sensor alerts in real time all over the information infrastructure. Data acquisition, storage and

processing should intrinsically follow a distributed structure, that is, every kind of data should have a processing corresponding to the data that is acquired with the monitored information infrastructure, which is important for system scalability.

All collected data are cleansed, normalized and stored in a distributed structure, which can already be used to support security information management and visualization. The data pre-processing mainly involves cleansing, normalization and collation. Data cleansing may include duplicate elimination, data calibration and filtering of the raw data from security sensors, such as IDS, firewall, network and system log records, security information and event management (SIEM), and NetFlow, etc.

At correlation and context management, data fusion and event processing and correlation take place. Data fusion is a technique to make overall sense of data from different sources which practically have different data structures. Data fusion is a technique to aggregate sets of evidence regarding a perceived situation. Dempster-Shafer evidence theory is a common technique for data fusion, which synthesizes belief levels of the individual data received from different sources so as to effectively reduce the false positive and false negative of security alerts. Furthermore, complex event processing, in which events are detected and correlated, may be utilized to exploit the higher level values out of the data.

At the end, security visualization is the transfer of organized data and information into meaningful patterns or sequence to be visualized. It is part of the comprehension layer of situational awareness. With all the data and events to create an integrated common picture, users can be prompted, immersed and informed by a common operational picture underpinned by CSAW. It is composed of vulnerability, assets, risks and instant status information. Such a consolidated cyber security picture allows decision-makers to make integrated risk analysis and corrective action planning.

CSAW undergoes a multi-level analysis framework which may be generally segmented into upstream and downstream [6], as illustrated in Figure 3.

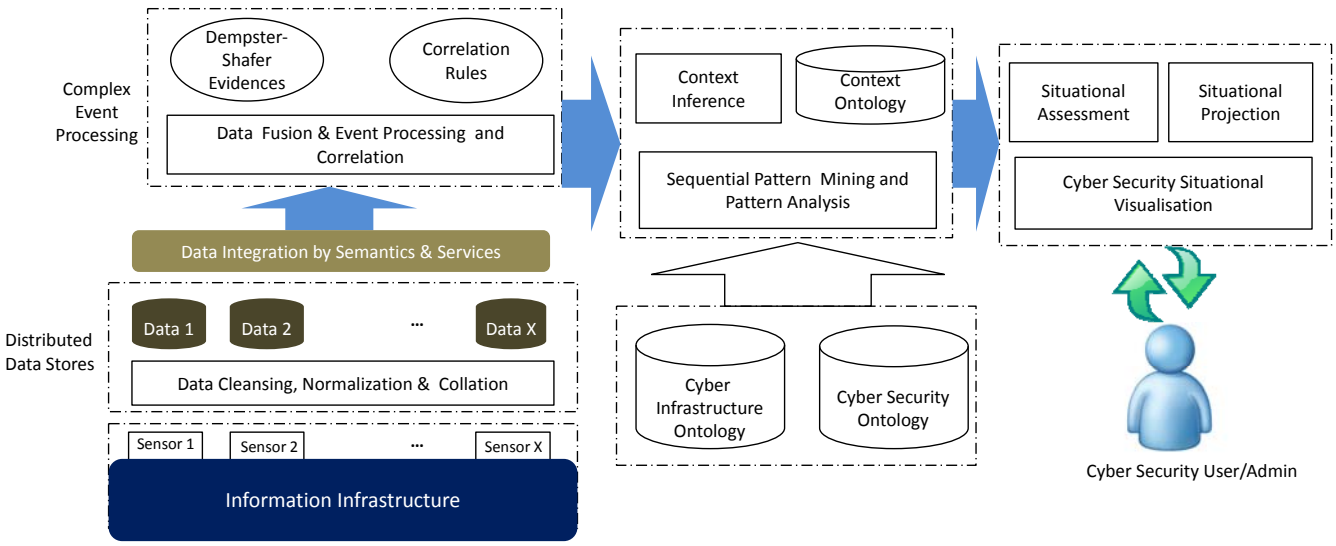

Figure 3. Upstream and downstream in the multi-level analysis framework of cyber security situational awareness

CSAW at the upstream is concerned with data pre-processing and the distributed data stores alongside constructing the information models, including data fusion based on Dempster-Shafer evidence theory for the delineation of cyber security situations.

At the start of the upstream is data pre-processing. Data pre-processing is designed to collect security data from different sensors, which are mostly embedded in cyber security toolkits, e.g., anti-virus, malware detection, IDS/IPS (intrusion detection/prevention systems), log management,

SIEM, etc. Pre-processing may include cleansing (e.g., duplicate elimination, data calibration, filtering/validation, etc.), normalization, collation, etc. Data validation mechanism is adopted to determine whether there is a successful attack. By comparing the conditions and the system configuration (e.g., operating system (OS) version, services running, etc.) necessary for a successful attack, non-impact attack alert could simply be removed. Finally, the security data will be normalized into a uniform format so as to be usable in the later stages of CSAW.

Data fusion is one of the advanced stages in the upstream, which may be carried out according to Dempster-Shafer evidence theory. It is worth noting that in the multi-level analysis framework of CSAW, data fusion, and event processing and correlation should be distinguished from the data cleansing (e.g., duplicate elimination, data calibration and filtering, etc.), normalization and collation. The former are aimed at exploiting higher level values out of the data, whereas the latter are to ensure the data is valid and correct.

CSAW at the downstream is concerned with the general processes of event processing and correlation analysis of various types of alert events from security sensors, sequential pattern mining and pattern analysis, and context inference and situational assessment and projection, and situational visualization.

For sequential pattern mining and pattern analysis, first, attack patterns are acquired through interactive knowledge discovery by applying frequent pattern mining algorithm, which helps discover the knowledge hidden in an event sequence. Then, the discovered frequent patterns and sequential patterns are transformed to the correlation rules of alert events. Finally, cyber security situation graph is dynamically generated.

Essentially, cyber security situational awareness (CSAW) constitutes a data driven framework of cyber defence. CSAW of an organization reflects the effectiveness of response to attacks since the ultimate goal of CSAW is to detect and ascertain advanced cyber-attacks.

3. Process of Data Mining Based Cyber-Attack Detection

Data mining based cyber-attack detection involves five general stages, as illustrated in Figure 4, that is, system monitoring and data capturing via various sensors, network/system/process logging and sniffing daemons/agents, and security devices, data pre-processing (e.g., cleansing, filtering, normalisation, etc.) at local data stores, event correlation and feature extraction (e.g., via big data processing, Hadoop Distributed File System (HDFS) and MapReduce), data mining (dimensionality reduction, classification, clustering) to detect misuse or anomaly, visualisations and interpretation of mining results.

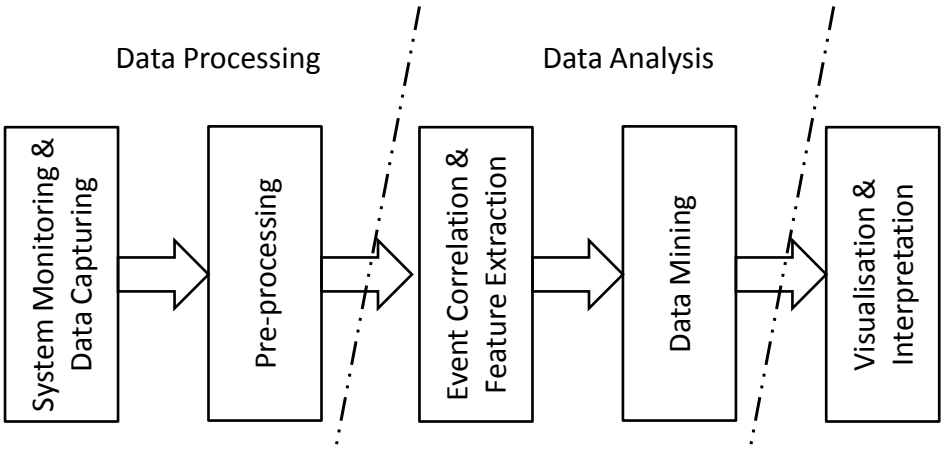

Figure 4. General stages in data mining based cyber-attack detection

Data mining based detection, when properly configured, has the capacity to become central nervous system of network. Data mining based detection can provide some useful derived functions, for

example, real-time monitoring and incident management for security related events which are collected from network, security devices, system, applications. It provides a workflow which helps to track and escalate the incident. It can also be used as log management, log consolidation, and generates reporting for compliance purpose. In other words, data mining based detection comes up with a comprehensive spectrum of respective components.

In a very summative manner, these stages may be put into three tiers, namely, processing, analysis and visualisation, with processing covering the first two stages and analysis the middle two stages.

3.1 Data Processing

Data Processing is the first tier of data mining based cyber-attack detection. The stage determines the format of the data. The format might be canonical or structured database.

Data Sources: data mining based detection utilises data feeds from various devices which include not only networking devices but also security devices. Data acquisition is responsible for collecting security log from security devices. It is collecting security events from various security devices, such as firewalls, sniffing daemons/agents in IDS/IPS, logging daemons/agents in SIEM, website protection devices, anti-virus. In addition, data acquisition collects attacks information and sensitive information from the mirror traffic of the switch.

Data store can be database and/or data warehouse systems. This layer is an interface to all data sources. It basically collects logs from various devices, normalizes the logs and stores in data stores. On all such data correlation rules are applied to get meaningful information.

In a way, the data source determines the types of attacks that can be detected. The two general categories are host-based detection and network-based detection.

For example, in network-based intrusion detection, network traffic is collected using sniffer software like tcpdump. Network-based intrusion detection employs network traffic as the main source of input. This involves placing a set of traffic sensors within the network. The sensors typically perform local analysis and detection and report suspicious events to a central location.

In host-based intrusion detection, data such as process activity, disk usage, memory usage and system calls are collected and commands such as netstat, ps and strace can serve this purpose. For host-based intrusion detection systems, the data source is collected from an individual host on the network. In particular, these systems employ their host's operating system audit trail as the main source of input. Because host-based systems directly monitor the host data files and operating system processes, they can determine exactly which host resources are the targets of a particular attack.

It is worth noting that network and host are just two examples of sources of security-related information.

The data collected by acquisition layer undergoes data pre-processing according to the requirements. Because of the differences in protocols used by logs from security devices and differences in formats of the internal packet payload, the identification and pre-processing by acquisition layer is required, that is standardization of security events.

Pre-processing involve a range of functions, e.g., cleansing, filtering, normalisation, etc., which may be performed at local data stores. Data cleansing is essential to remove the noise and eliminate components with missing data values. Normalization happens in two ways. It first normalizes the values such as time zone, priority, severity in to common format, and then the data structure in to common format. Sometime pre-processing carries out aggregation. Filtering increases efficiency and accuracy and reduce processing time.

Security monitoring is relatively a broader concept. The basic functions of security monitoring may include security data acquisition, security event correlation, security status overview, and security analysis, for instance, monitoring the network border real-time and making deep analysis; integrating various security events and correlating them; displaying the whole network security posture real-time and forming several types of security analysis reports.

3.2 Data Analysis

This is heart of data mining based cyber-attack detection. The massive data collected by acquisition layer is for centralized storage and analysis in analysis tier, so as to extract the major information concerned.

The collected data is analysed in this step to determine whether the data is anomalous or not. Here, it may involve feature selection and threat correlation. Feature selection generates from large dataset feature vectors. Threat correlation uses the artificial intelligence to sort through multiple logs and log entries to identify attackers.

Analysis tier realizes functions of real-time analysis and deep analysis. Among them, the real-time analysis includes event positioning, time correlation analysis, knowledge base, security situation analysis, alarm generation and storage indexing; while deep analysis includes data storage and statistics, artificial intelligence based analysis and reporting.

Then, various data mining techniques are used to retrieve data from database. Some transformation routine can be performed here to transform data into desired format. Then data is processed using various data mining algorithms. Major step of this stage is to classify data. So, data is organized in some pattern. The classification depends on the analysis schemata being used.

3.3 Response

Obviously after all, actions need to be taken in response to the detected attacks. Response can be set to be performed automatically. In case of manual analysis situations it can be done manually, which means the final ascertainment by human administrators for determination about cyber-attacks and mitigation and decision.

Detection system alerts the system administrator that an attack has happened using several methods like e-mail, alarm icons and visualization techniques. Detection system can also stop or control attack by closing network ports or killing processes.

Response also takes the form of refinement: This does tuning task based on previously detected attacks, and updating and re-training of classifiers based on newly detected cyber-attacks. It can reduce more false positive levels. It supports to have more security tools which ensure that the raised alert is valid or not.

A more comprehensive approach for monitoring a myriad of diverse heterogeneous event sources for cyber-attack detection can yield a better situational awareness of the threats in cyberspace, and thus improve detection accuracy and minimize false alarms by correlating security events among these diverse sources.

3.4 Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of random variables under consideration. Two techniques for dimensionality reduction are feature extraction and feature selection. [7] presents a comparative numeric study on dimensionality reduction algorithms in conjunction with detection techniques under misuse and anomaly detection models.

Feature extraction refers to the mapping of the original high-dimensional data onto a lower-dimensional space. Through feature extraction, all the original features are combined into a new reduced set of features. Examples of feature extraction algorithms are Principal Component Analysis (PCA) and Independent Component Analysis (ICA).

Feature selection is a process to find the most important and optimal subset of features for building powerful learning models.

In other words, feature selection identifies the most relevant features/attributes that only are to be used. The search process that finds an optimal subset of features can be undertaken by using meta-heuristic algorithms, e.g., Genetic Algorithm (GA) and Particle Swarm Optimization (PSO).

The collected data is substantially large and cannot be used as it is, so subset of this data is selected by creating feature vectors that contain only necessary information needed for cyber-attack detection. In network based intrusion detection, it can be IP packet header information which includes source and destination IP addresses, packet length, layer four protocol type and other flags. In host-based intrusion detection it includes user name, login time and date, duration of session and number of opened files.

An efficient feature selection method can eliminate irrelevant and redundant data; hence it can significantly improve classification processing times. In some cases, it can even improve classification accuracy by removing misleading noise.

However, care should be exercised in how feature selection is applied especially for real-world application in terms of relevancy and efficiency. A smaller number of relevant features will improve classification processing times from an efficiency perspective. This comes at the cost of the computational process of performing feature selection, which in some cases can be time-consuming.

Applying feature selection to a data set where its attributes can change rapidly and diversely might not be able to generate feature sets in close to real-time. Due to the inherently dynamic nature of cyber security, in terms of feature selection, yesterday's selected features from a static data set might not be relevant for tomorrow's dynamically different data set. A new attack class can make different features important, and different feature sets may or may not be relevant even at the millisecond scale. Thus, it is important to rethink the relevancy of feature sets from data sets that are out of date, very static, or lack of diversity.

4. Detection Models and Learning Modes

According to whether attack's patterns are known or unknown, there are two detection models [8], i.e., misuse detection and anomaly detection.

4.1 Misuse detection

Also known as signature based detection, misuse detection uses patterns of well-known attacks to identify attacks. A signature can be viewed as a pattern representing a well-known attack or threat. Pattern of known malicious activity is stored as signatures. Misuse detection works by searching for the traces or patterns of well-known attacks in the signature base. If a pattern match is found, it signals an event then an alarm is raised.

To perform misuse detection method, each scenario needs to be described or modelled. Misuse detection is based on extensive knowledge of patterns associated with known attacks provided by human experts. Also the signature base needs to be updated when any new intrusion is detected.

As derived from its principle, misuse detection is good at detecting known attacks. The main disadvantage of misuse detection is that it can only detect attacks that follow pre-defined patterns but it is unable to detect any future (unknown) attacks that do not have any matched pattern stored in the system.

4.2 Anomaly Detection

Also known as behaviour-based detection, or profile-based detection, anomaly detection pre-stores a baseline of profiles about system's usual behaviour and normal data in network, e.g., load on network traffic, protocol and packet size etc. An anomaly can be viewed as a variation or deviation to a known behaviour.

Anomaly detection attempts to determine whether deviation from an established normal behaviour profile can be flagged as an anomaly. Anomaly detection consists of pre-establishing the normal behaviour profiles for users, programs, or other resources of interest in a system, and observing the actual activities as reported in the audit data to ultimately detect any significant deviations from these profiles.

Anomaly detection monitors new instances. The new instances are compared with the profiles stored in the baseline store. If any kind of divergence is found, it is said an anomaly has detected.

Generally, normal usage patterns can be established using statistical measures on system audit data and network data. The baseline of usual behaviour is defined by system administrator, and a profile represents expected or normal behaviour obtained by monitoring regular activities, connections in a network, and no. of hosts, network routers and users for a period of time. Profiles can be categorized as either dynamic or static.

Anomaly detection is about discrimination of legitimate patterns of activities characterizing system normality.

Strength of anomaly detection is its ability to detect previously unknown threats and attacks (which do not have signatures or labelled data corresponding to them) as deviations from normal usage. Moreover, unlike misuse detection (which builds classification models using labelled data and then classifies an observation as normal or attack), anomaly detection does not require an explicitly labelled training data set, which is very desirable, as labelled data is difficult to obtain in a real network setting.

Anomaly detection is a great deal used in behavioural analysis and other forms of analysis in order to assist in knowledge concerning the detection, recognition and forecast of the occurrence of these anomalies or attacks. Anomaly detection can also term as outlier detection. Based on knowledge of normal behaviour it looks for anomalous behaviour or deviations from the recognized baseline.

Anomaly detection's most obvious drawback is its high false positive [9]. Any significant deviations from the profile are then reported as possible attacks. Such deviations are not necessarily actual attacks. They may simply be new network behaviour that needs to be added to the profile. In addition, selection of the correct set of system features to measure is ad hoc and based on experience, which is quite difficult to perform.

4.3 Learning Modes

A machine learning approach usually consists of two phases: training and testing. Often, the following steps are performed:

(i) Identify attributes (features) and classes from training data;
(ii) Identify a subset of the attributes necessary for classification (i.e., dimensionality reduction);
(iii) Train the model using training data;
(iv) Apply the trained model to classify the test data.

In the case of misuse detection, in the training phase each misuse class is learned by using appropriate samples from the training set. In the testing phase, new observed data is run through the model and is classified as to whether it belongs to one of the misuse classes. If the observed data does not belong to any of the misuse classes, it is classified as normal.

In the case of anomaly detection, the normal traffic pattern is defined in the training phase. In the testing phase, the learned model is applied to new observed data. If the observed data belong to any of the profile classes, it is classified as normal, which means things look as expected. Otherwise, if the observed data does not belong to any of the profile classes, it is classified as anomalous, which means that something unexpected appears.

Determining complex attacks often requires observing the anomalies and the correlations between them to uncover scenarios or attack plans. To achieve this goal, alert correlation has been studied to reduce the volume of alerts generated by the system. Clustering, similarity measures, or expert knowledge is used to correlate anomalies and then reveal the complex patterns of attack scenarios.

There are three main types of machine learning/data mining approaches: unsupervised, semi-supervised, and supervised. In unsupervised learning problems, the main task is to find patterns,

structures, or knowledge in unlabelled data. When a portion of the data is labelled during acquisition of the data or by human experts, the problem is called semi-supervised learning. The addition of the labelled data greatly helps to solve the problem. If the data are completely labelled, the problem is called supervised learning and generally the task is to find a function or model that explains the data. The approaches such as curve fitting or machine-learning methods are used to model the data to the underlying problem. The label is generally the variable which experts assume has relation to the collected data.

4.4 Data Mining Based Detection Architecture

Misuse detection and anomaly detection themselves are about the how attacks can be perceived from different perspectives, but not about when the detection is being executed and what specific algorithms are used for calculating the detection. As such data mining can literally be applied in either misuse detection or anomaly detection, e.g., data mining may be run on audit data off-line to find patterns for misuse detection. As for anomaly detection, since its aim is to detect unknown attacks, data mining seems to just fit the purpose, though what is actually executed in anomaly detection is to first form the normal baselines, i.e., the profiles.

For example, data mining is applied for misuse detection systems, JAM (Java Agents for Meta-learning) [10] [11], MADAM ID (Mining Audit Data for Automated Models for Intrusion Detection) [10] [11], Minnesota Intrusion Detection System (MINDS) [12]. At the same time, data mining is also applied to anomaly detection systems, ADAM (Audit Data Analysis and Mining) [13], IDDM (Intrusion Detection using Data Mining) [14].

A clearer view may be to look at when detection is being executed, i.e., off-line versus run-time on-line. In an on-line detection system, suppose there is already a signature base with which any known attacks can be detected, then what remains to be done is that, based on the stream data captured from the computer system, the detection system should work to spot whether there is presence of unknown attacks. Once any unknown attack has been spotted and ascertained, then it can be added into the signature base for future use. This actually shows how the two loops of detection intertwine with one another in an on-going basis.

In practice, to be a holistic solution of cyber-attack detection, it is necessary to handle seamlessly both known attack patterns and unknown attacks as well. Therefore, both supervised machine learning and data maiming, namely misuse detection and anomaly detection need to be integrated. In a way, misuse detection works in a reflexive loop based on pattern matching using the attack's signature base, while data mining based cyber-attack detection works in a deliberate loop, as illustrated in Figure 5.

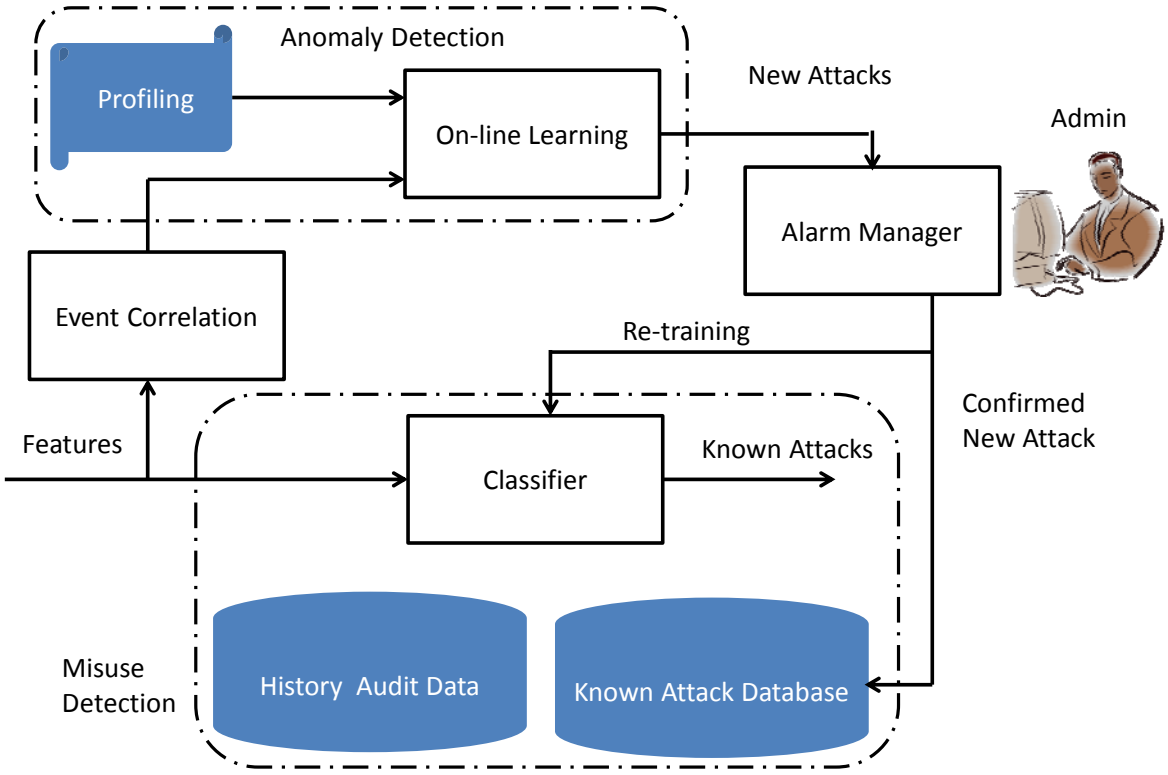

Figure 5. Multi-loop learning for cyber-attack detection

The Misuse Detection loop analyses the features which supposedly have undergone data pre-processing, to detect the network connections that correspond to attacks for which signatures are available, and then to remove them from further analysis. The remaining features are fed into an Anomaly Detection loop, which uses the on-line machine learning to detect the abnormal behaviours. New attacks confirmed by administrator are updated to the attack database and also re-train the classifier.

While the anomaly detection loop aims to detect actual attacks and other abnormal activities in the network traffic, the profiling module detects the dominant modes of traffic to provide an effective profile of the network to the analyst.

In the off-line detection, audit data is used to train the attack database and also to establish profiles. In the on-line detection phase:

(i) Features after the pre-processing are inputted as test data to the trained classifier which classifies the features in active database to detect known attacks.

(ii) Once an unknown attack is detected by the on-line machine learning algorithm along with the event correlation, and further confirmed by the administrator, the features associated with to the new attack are updated to the attack database and used to re-train the classifier.

(iii) In practice, logs are distributed across the information infrastructure. There is a need for a distributed file structure that maintains the various distributed local logs and manages them as an integrated file system. Such a distributed file structure handles all the logs sent from the local log stores.

The alarm manager is prepared to prompt the administrator when an attack is detected by the on-line machine learning. If an alarm is prompted, the alarm manager asks for a response of confirmation. There are three types for confirming: an attack, a false alarm or an unclear event. The alarm manager will record the response to a database and invoke the machine learning mechanism to learn about the response.

It goes without saying, there needs to be an interface that provides the administrator with control on the data mining based detection system.

5. Data Mining Techniques for Cyber-Attack Detection

Different data mining techniques such as classification, clustering, and association rules are used for detecting cyber-attacks, either run on system audit data or network data.

5.1 Classification

Classification is the process of assigning data items to pre-defined classes. [15] The result of this process will be a classifier based on association rules or decision trees. For example, suppose sufficient “normal” and “abnormal” audit data is gathered for a user or a program, then a classification algorithm is applied to learn a classifier that can label or predict new audit data as belonging to the normal class or the abnormal class.

Classification categorizes the datasets into pre-defined classes. There are two steps, i.e., training and prediction. In the first step, classifier is trained by analysing a training set made up of data instances and their associated class labels. Because the class label of each training instance is provided, this is known as supervised learning. In the second step, the trained classifier is used to predict the class for unlabelled data instance.

An algorithm that implements classification is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category. The classes are pre-defined in training phase. In terms of the pre-defined classes, classification may have two cases – Binary and Multiclass classification. in binary classification, only two classes are involved, whereas multiclass classification involves assigning an instance of dataset to one of several classes [16].

In principle, classification may be used for both misuse detection and anomaly detection, but mostly used for misuse detection. As for misuse detection, the detection can be formulated as a classification problem. Suppose sufficient audit data has been gathered in which each data instance will be labelled as either “normal” or “abnormal”. We then use classification algorithm on audit data to train a classifier. When run on new observed audit data, this classifier will then predict the class of the new data instance as “normal” or “abnormal”.

Common classification techniques include decision tree, Naive Bayes classifier, K-nearest neighbour classifier, support vector machine, as well as neural network, genetic algorithm and fuzzy logic [17].

5.1.1 K-Nearest Neighbour

The k-Nearest Neighbour (k-NN) algorithm is a simple classification algorithm. [18] **[19]** all available cases (training data) are stored and new cases (test data) are classified based on a similarity measure (e.g., Euclidean distance) in the feature space. In the training data, the feature vectors are associated with respective class labels.

For a new observed data, k-NN computes the distances between the new case (an unlabelled vector) and all existing cases (labelled vectors) and then the test data is allotted to the class most common amongst its k nearest neighbours. It can be understood that the new observed data is categorized by the majority vote of its neighbours. If k=1, then the new data, without any doubt, is assigned to the class of its nearest neighbour. Large prediction time is needed if the value of k is large [19].

5.1.2 Decision Trees

Decision tree is a recursive and tree-like structure for expressing classification rules. [19] It uses divide and conquer method for splitting of the data according to attribute values. The splitting process repeats for every child node till all selected attributes are considered. Decision tree converts

the given dataset in to a tree structure. The nodes of the tree represent the features and the edges represent the association between the features by value of features. Each leaf node, i.e., the lowest level of the node represents the class label.

A decision tree partitions the data with same properties into groups and these groups are kept as analogous as possible. Decision tree is supervised learning. It takes a set of classified data as input, executes the algorithm on that and provides a tree as output where each leaf can be expressed as a decision and each intermediate node epitomizes a test. [20]

To classify a data item, it proceeds from root node to leaf node. That is, one starts at the root of the decision tree and follows the branch indicated by the outcome of each test until a leaf node is reached. The name of the class at the leaf node is the class of an unknown data item. The best attribute to divide the subset at each stage is selected using the information gain of the attributes.

5.1.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is supervised learning for classification and prediction. SVM is a classifier based on finding a separating hyperplane in the feature space between two classes in such a way that the distance between the hyperplane and the closest data points of each class is maximized. It separates data points into two classes +1 and −1 using hyperplane. +1 represents normal data and −1 for suspicious data. [21]

SVM maps the input vector into a higher dimensional feature space and obtains the optimal separating hyperplane in the higher dimensional feature space. When the two classes are not separable, slack variables are added and a cost parameter is assigned for the overlapping data points.

SVMs are well known for their generalization ability and are particularly useful when the number of features, m, is high and the number of data points, n, is low ($m >> n$). SVM is moderately insensible to the number of data points and the classification complexity doesn't depend on the dimensionality of the feature space. So, SVM can potentially learn a large set of patterns.

Various types of dividing classification surfaces can be realized by applying a kernel, such as linear, polynomial, Gaussian Radial Basis Function (RBF), or hyperbolic tangent. SVMs are binary classifiers and multi-class classification is realized by developing an SVM for each pair of classes.

5.2 Clustering

Clustering is the process of seeking to describe the available datasets by grouping them into common clusters or categories. Clustering could be utilized to detect anomaly. Unlike misuse detection which requires labelled data set and supervised algorithms, anomaly detection works on unlabelled dataset and uses unsupervised clustering algorithms.

Clustering is a widely used data mining technique which groups similar items, to obtain meaningful groups/clusters of data items in a data set. These clusters represent the dominant modes of behaviour of the data items determined using a similarity measure. [22]

The main advantage of clustering algorithm is the ability to learn from and detect attacks in the audit data without explicit descriptions of various attack classes (signatures) which need to be provided by security administrators.

There are several clustering models. In connectivity models (e.g., hierarchical clustering), data points are grouped by the distances between them. In centroid models (e.g., k-means), each cluster is represented by its mean vector. In distribution models (e.g., Expectation Maximization algorithm), the groups are assumed to be acquiescent to a statistical distribution. Density models group the data points as dense and connected regions (e.g., Density-Based Spatial Clustering of Applications with Noise (DBSCAN)). Lastly, graph models (e.g., clique) define each cluster as a set of connected nodes (data points) where each node has an edge to at least one other node in the set.

5.2.1 k-means

k-means is a well-known and widely used clustering algorithm. k-means can be used to automatically recognize groups of similar instances/items/points.

k-means clustering is to allocate ‘m’ data items into ‘k’ clusters and uses Euclidean distance metric for measuring the similarity. [23], The algorithm classifies instances to a pre-defined number of clusters specified by the user (e.g., assume k clusters).

The first step is to choose a set of k instances as preliminary centroids (centres of the clusters). To achieve this, k data items are selected arbitrarily from dataset, one for each cluster, and usually as far as possible from each other. Then, each data item is assigned to a cluster by computing the distance (e.g., Euclidean distance) between the data item and each cluster centroid and allocating it to the nearest cluster. In this way, the intra-cluster similarity remains high and inter-cluster similarity goes low [24]. Next, the averages of all clusters is recalculated and renewed and the afresh computed mean is assigned as the new centroid. This process is replicated until the criterion function has converged.

5.2.2 k-medoids clustering algorithm

k-medoids is clustering by partitioning algorithm as like as k-means algorithm. The most centrally situated instance in a cluster is considered as centroid in place of taking mean value of the data items in k-means clustering. This centrally located data item is called reference point, i.e., the medoid. A medoid is a data point which acts as an exemplar for all other data points in the cluster. attempts to minimize the distance between data points and its centre (centroid). [25]

The k-means algorithm is very sensitive to outliers because if there is a data item with a very large value, the data distribution may be biased or distorted [26]. In this case, k-medoids is more robust in presence of noise and outliers because in k-medoids clustering algorithm, the partitioning method is performed based on the principle of minimizing the sum of dissimilarities between data items in a cluster and thus medoid is less influenced by outliers. [26].

5.2.3 Expectation Maximization (EM) Clustering

Expectation Maximization (EM) clustering is a variant of k-means clustering and is widely used for density estimation of data points in an unsupervised learning. In the EM clustering, the algorithm tries to find the parameters which maximize the likelihood of the data, assuming that the data is generated from k normal distributions. The algorithm learns both the means and the covariance of the normal distributions. This method requires several inputs which are the data set, the total number of clusters, the maximum error tolerance and the maximum number of iteration.

The algorithm is recursive in nature. The EM clustering undertakes two steps, namely Expectation (E-step) and Maximization (M-step). The goal of E-step is to calculate the expectation of the likelihood (the cluster probabilities) for each instance in the dataset and then re-label the instances based on their probability estimations. The M-step is used to re-estimate the parameter values from the E-step results. The outputs of M-step (the parameter values) are then used as inputs for the following E-step. These two processes are performed iteratively until the convergence is reached. [27].

5.3 Association Rule Mining

The concept of association rule mining is to find all co-occurrence relationships called associations. It finds frequent sets of items (e.g., combinations of items that are purchased together in at least N transactions in the database), and from the frequent items sets such as {X, Y}, generates association rules of the form: $X \rightarrow Y$ and/or $Y \rightarrow X$. In general, association rule mining is considered as an unsupervised technique.

The goal of association rule mining is to discover previously unknown association rules from the data. An association rule describes a relationship among different attributes: IF (A AND B)

THEN C. This rule describes the relationship that when A and B are present, C is present as well. Association rules have metrics that tell how often a given relationship occurs in the data. The support is the prior probability (of A, B, and C), and the confidence is the conditional probability of C given A and B.

Association rule mining is the process of finding all the association rules that pass the condition of minimum support and minimum confidence. In order to mine these rules, the support and confidence values have to be computed for all of the rules and then compare them with the threshold values to prune the rules with low values of either support or confidence.

A limitation of traditional association rule mining is that it only works on binary data (i.e., an item was either purchased in a transaction (1) or not (0)). In many real-world applications, data are either categorical (e.g., IP name, type of public health intervention) or numerical (e.g., duration, number of failed logins, temperature). For numerical and categorical attributes, Boolean rules are unsatisfactory.

6. Outlook

There is no doubt that cyber-attack detection is big data processing and analytics problem [28]. Data driven framework makes a systematic approach to addressing cyber-attack detection issues. In the latitudinal perspective, big data provides opportunities of utilising data sources of heterogeneous types and formats. In terms of depth of intelligence, that is what data mining is expected of.

As an area in itself, data mining has a fairly established set of models and techniques. However, facing advanced cyber-attack problems, one of the technical challenges lying ahead is how the models of supervised and unsupervised learning, the models of cyber-attack detection, and the off-line and on-line system modes should be interwoven all together into an organic robust cyber-attack detection system.

References

[1] S. Noel, D. Wijesekera and C. Youman (2002). Modern intrusion detection, data mining, and degrees of attack guilt. In: D. Barbará and S. Jajodia (eds.), Applications of Data Mining in Computer Security, series Advances in Information Security, Volume 6, Springer Science+Business Media New York, 2002, pp. 1-31.

[2] W. Lee and S. J. Stolfo (2000). A framework for constructing features and models for intrusion detection systems. Information and System Security, Vol. 3, No. 4, pp. 227-261.

[3] A. L. Buczak and E. Guven (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, Vol. 18, No. 2, Second Quarter 2016 1153-1176

[4] I. Brahmi, et al. (2012). Towards a multiagent-based distributed intrusion detection system using data mining approaches. In: L. Cao, et al. (eds), Agents and Data Mining Interaction (ADMI'2011), Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, Vol 7103, pp. 173-194.

[5] G. P. Tadda and J. S. Salerno (2010). Overview of cyber situation awareness. In: S. Jajodia et al, (eds), Cyber Situational Awareness-Issues and Research, vol. 46, Springer, 2010, pp. 15-35.

[6] H. Tianfield (2016). Cyber security situational awareness. In: Proceedings of 2016 IEEE International Conference on Smart Data (SmartData'2016), Chengdu, China, 15-18 December 2016, pp. 782-787

[7] I. Syarif, A. Prugel-Bennett and G. Wills (2012). Data mining approaches for network intrusion detection: from dimensionality reduction to misuse and anomaly detection. Journal of Information Technology Review, Vol. 3 No. 2, May 2012, pp. 70-83

[8] M. K. Asif, et al. (2013). Network Intrusion detection and its strategic importance. In: 2013 IEEE Business Engineering and Industrial Applications Colloquium (BEIAC'2013), Langkawi, Malaysia, 7-9 April 2013, pp. 140-145

[9] M. Panda, A. Abraham, M. R. Patra (2012). A hybrid intelligent approach for network intrusion detection. International Conference on Communication Technology and System Design 2011, Procedia Engineering, Vol. 30, 2012, pp. 1-9

[10] W. Lee, S. Stolfo and K. Mok (2000). Adaptive intrusion detection: a data mining approach. Artificial Intelligence Review, vol. 14, no. 6, pp. 533-567.

[11] W. Lee, S. J. Stolfo and K. W. Mok (2002). Algorithms for mining system audit data. In: T. Y. Lin, et al. (eds.), Data Mining, Rough Sets and Granular Computing, Volume 95 of the series Studies in Fuzziness and Soft Computing, Springer-Verlag Berlin Heidelberg, 2002, pp 166-189

[12] L. Ertöz, et al. (2004). MINDS minnesota intrusion detection system. In: H. Kargupta, et al. (eds.), Data Mining Next Generation Challenges and Future Directions. AAAI Press, 19 November 2004, 528 pages

[13] D. Barbara, N. Wu and S. Jajodia (2001). Detecting novel network intrusions using Bayes estimators. In: Proceedings of the 2001 SIAM International Conference on Data Mining. http://epubs.siam.org/doi/pdf/10.1137/1.9781611972719.28

[14] T. Abraham (2001). IDDM: Intrusion detection using data mining techniques. Technical Report DSTO-GD-0286, DSTO Electronics and Surveillance Research Laboratory. http://dspace.dsto.defence.gov.au/dspace/bitstream/1947/3750/1/DSTO-GD-0286%20PR.pdf

[15] D. K. Denatious and A. John (2012). Survey on data mining techniques to enhance intrusion detection. International Conference on Computer Communication and Informatics (ICCCI'2012), Coimbatore, India, Jan. 10-12, 2012, 5 pages

[16] S. Har-Peled, D. Roth, D.Zimak (2003). Constraint classification for multiclass classification and ranking. In: Becker, B., Thrun, S., Obermayer, K. (Eds) Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference on Neural Information Processing Systems (NIPS'2002), December 9-14, 2002, Vancouver, Canada, MIT Press, pp. 809-816

[17] J. Han, M. Kamber and J. Pei (2011). Data mining: concepts and techniques. 3rd edition, Morgan Kaufmann, 25 July 2011, 744 pages

[18] M.A. Maloof (2006). Machine Learning and Data Mining for Computer Security, Springer-Verlag, 2006.

[19] H. Tribak, et al. (2012). Statistical analysis of different artificial intelligent techniques applied to intrusion detection system. In: 2012 International Conference on Multimedia Computing and Systems (ICMCS'2012), Tangier, Morocco, 10-12 May 2012, 7 pages

[20] G. L. Agrawal and H. Gupta (2013). Optimization of C4.5 decision tree algorithm for data mining application. Int. J. of Emerging Technology and Advanced Engineering, Vol. 3, No. 3, March 2013, pp. 341-345

[21] M. Ektefa, et al. (2010). Intrusion detection using data mining techniques. In: 2010 International Conference on Information Retrieval & Knowledge Management (CAMP'2010), Shah Alam, Selangor, Malaysia, 17-18 March 2010, pp. 200-203

[22] L. Portnoy, E. Eskin and S. Stolfo (2001). Intrusion detection with unlabeled data using clustering. In: Proceedings of the ACM Workshop on Data Mining Applied to Security (DMSA'2001), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.126.2131

[23] H. Om and A. Kundu (2012). A hybrid system for reducing the false alarm rate of anomaly intrusion detection system. 1st Int'l Conf. on Recent Advances in Information Technology (RAIT'2012), Dhanbad, India, 15-17 March 2012, 6 pages

[24] T. Kanungo and D. M. Mount (2002). An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, Jul 2002, pp. 881-892

[25] R. Chitrakar and H. Chuanhe (2012). Anomaly based intrusion detection using hybrid learning approach of combining k-medoids clustering and Naïve Bayes classification. In: 2012 8th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM'2012), Shanghai, China, 21-23 Sept. 2012, 5 pages

[26] T. Velmurugan and T. Santhanam (2010). Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points. Journal of Computer Science, vol. 6, no. 3, 2010, pp. 363-368.

[27] D. Kumar and Suman (2011). Performance analysis of various data mining algorithms: A review. Int. J. of Computer Applications, Vol 32, No.6, October 2011, 7 pages

[28] A. Razaq, H. Tianfield and P. Barrie (2016). A big data analytics based approach to anomaly detection. In: Proceedings of 2016 IEEE/ACM 3rd International Conference on Big Data Computing Applications and Technologies (BDCAT'2016), Shanghai, China, 6-9 December 2016, pp. 187-193.